West Nile Virus Prediction

Project 4 : Predict West Nile virus in mosquitoes across the city of Chicago

Group Members : Joe, Mild, Gear

Table of Contents

The Virus

Most commonly spread to humans through infected mosquitos, symptoms ranging from a persistent fever, to serious neurological illnesses and death

It is believed that hot and dry conditions are more favorable for West Nile virus than cold and wet.

The Situation

The West Nile Virus have infected over 4,200 people and killed 177 people in 2006.

The CDC wants us to explore weather, location, testing, and spraying data, to predict when and where different species of mosquitoes will test positive for West Nile virus.

Thus, allowing them to effectively allocate resources towards preventing transmission

01

EDA & Cleaning Data

The Data

Train Dataset

12 columns with features: location, no of mosquitos, and virus positive/negative

Test Dataset

11 columns of data like train dataset with no virus present result

The Weather

22 columns with features: temperature, humidity, and station location

Spray Data

4 columns with spray location, date and time

The Weather Station

Station 1

Chicago O’hare Intl Airport

- Lat: 41.995 Lon: -87.933
- Elev: 662 ft. above sea level

Station 2

Chicago Midway Intl Airport

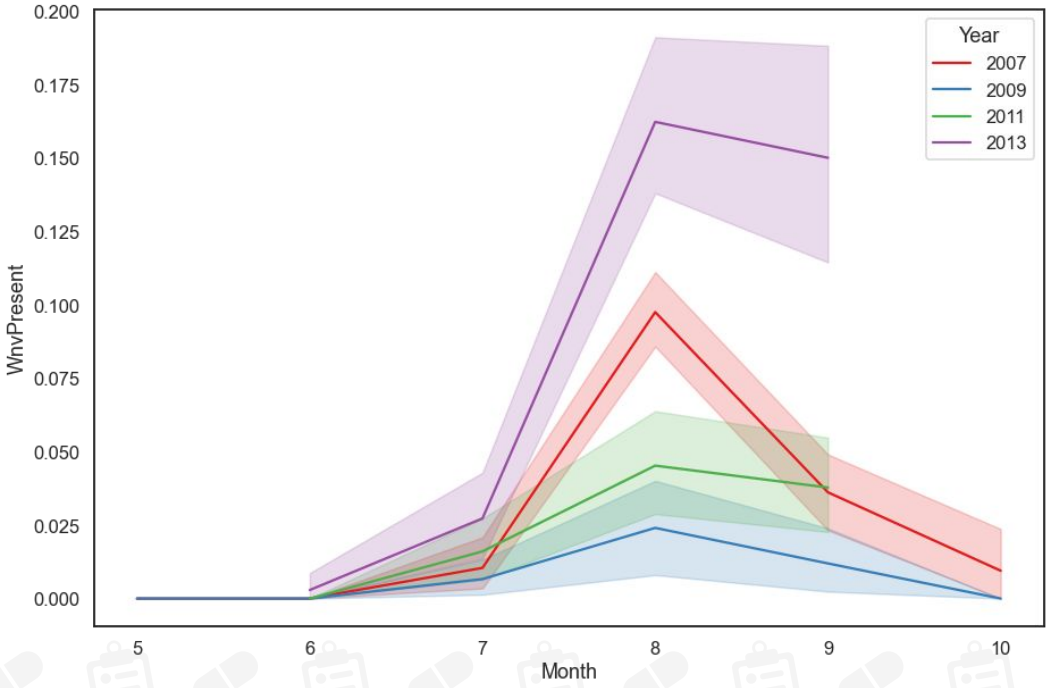
- Lat: 41.786 Lon: -87.752
- Elev: 612 ft. above sea level

Data at a Glance

Aug

Peaked Positive

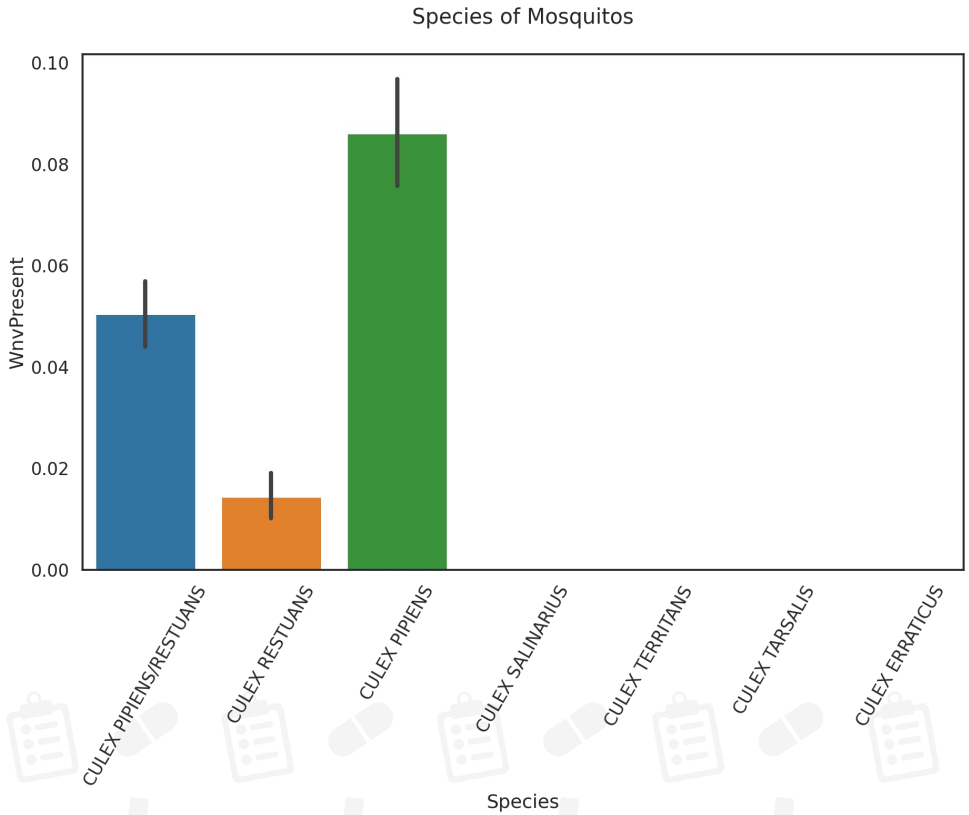
Number of patients tested positive for West Nile Virus



Data at a Glance

3 Species

There are 3 species that have a West Nile virus

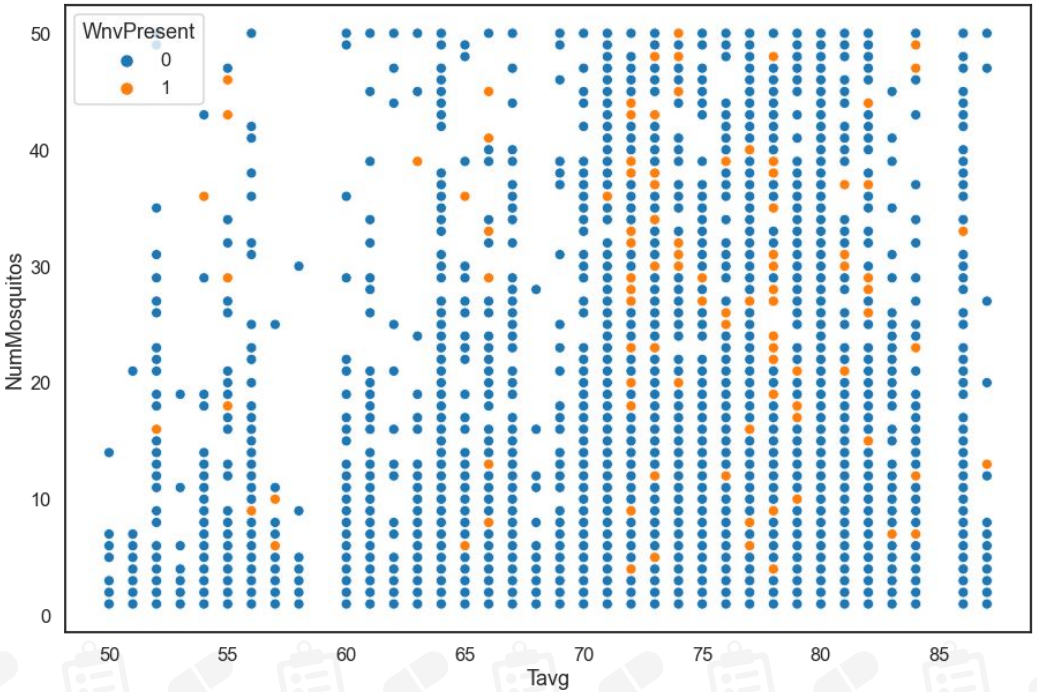


Data at a Glance

>72 f

High Temp

Positive correlation between temperature and positive virus result

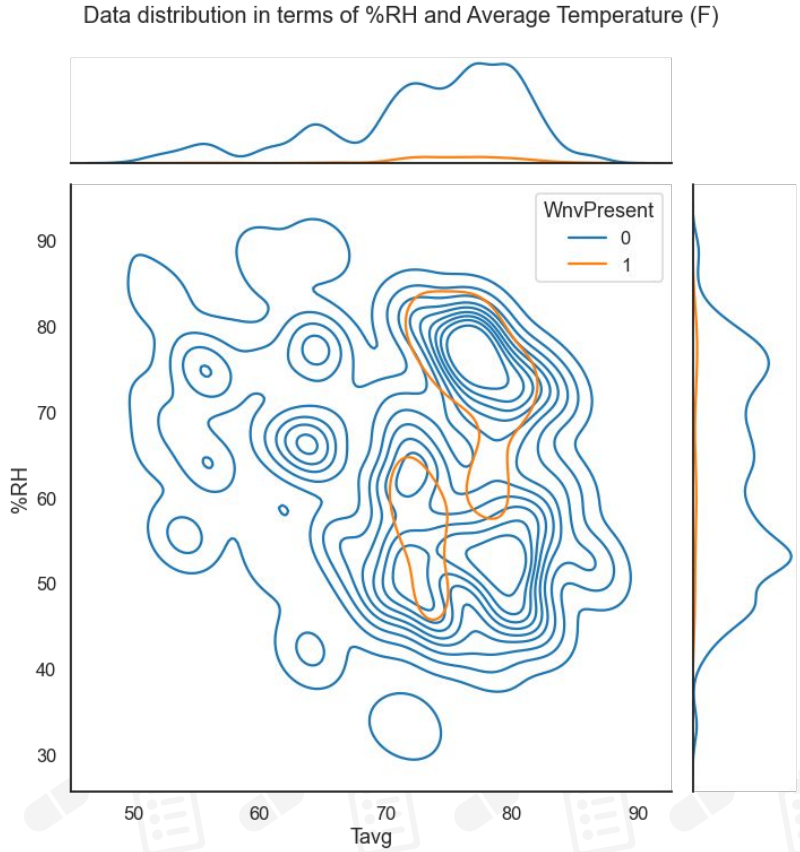


Data at a Glance

>50 %RH

%RH
(Relative Humidity)

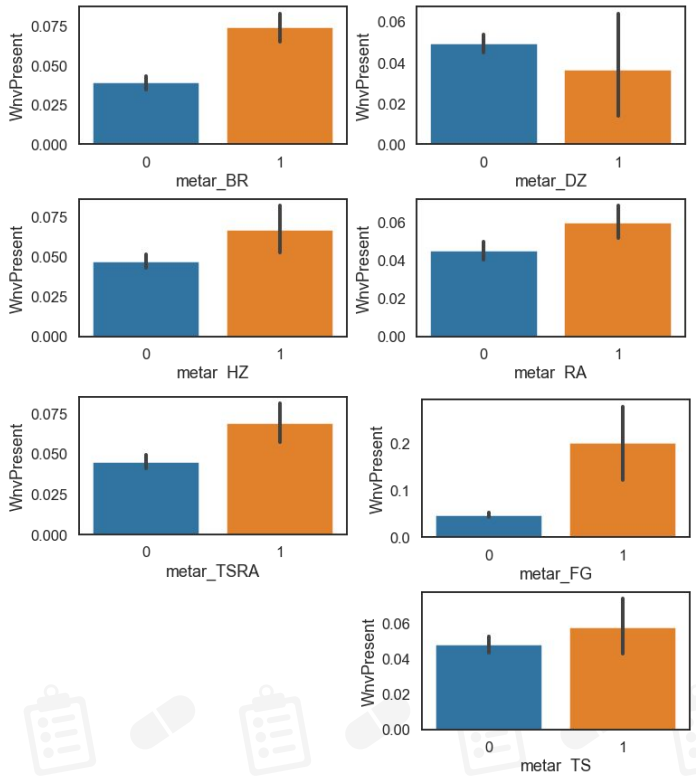
Virus spread widely in Hot and Humid condition



Data at a Glance

Weather Types:

- BR - Mist
- DZ - Drizzle
- HZ - Haze
- RA - Rain
- TSRA - Thunder Storm + Rain
- FG - Fog
- TS - Thunder Storm



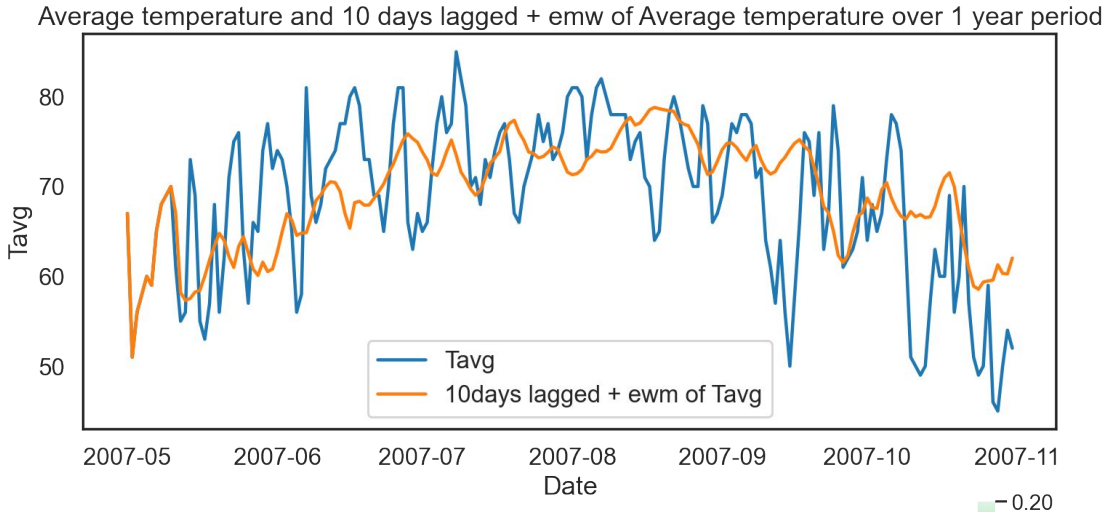
Data at a Glance

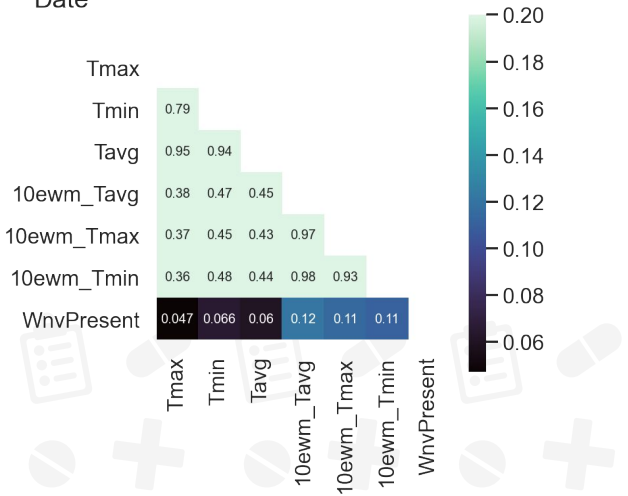
x2

Correlation

By adding
10 days lagged

Stronger positive correlation of 10 days lagged weather data





From the Data

Dropped Columns

Dropped columns such as traps address, rain depth, water volumes due to lack of data

Distance Calculations

Converted the latitude and longitude data into km to calculate distance to weather station

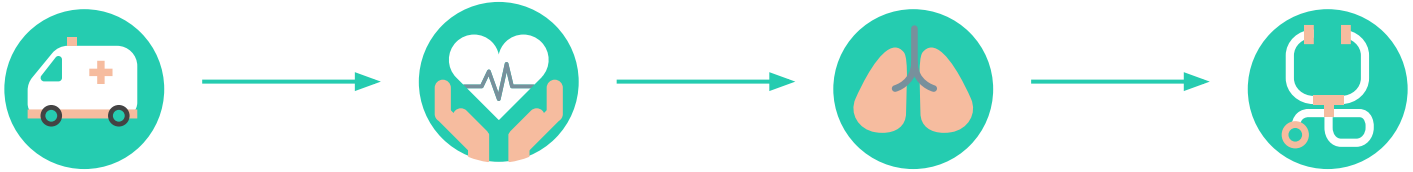
Station Mapping

Matching meteorological data from the nearest station to each trap

02

Features Engineering & Modeling

Preparing the Data



EDA

Clean, cluster and one hot encoded the data

SMOTE

Fixing the imbalance data by increasing minority class

PCA

Reduce the dimensionality of large data sets by grouping them

Models

Test models to predict when the virus will present

Features Engineering

%RH

Relative Humidity calculated from average temperature minus dew point

Date-Time Clustering

Group the dates data into months and weekly clustering for each year

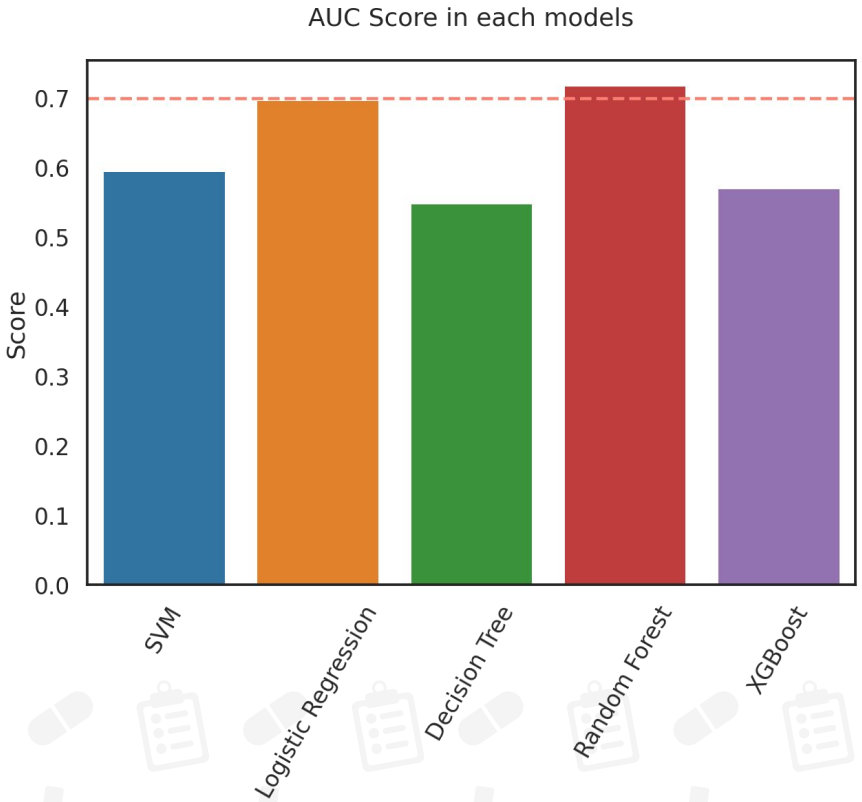
Location Clustering

Used KMeans to cluster the lat and long data to group location data

Modeling and Predictions

From the features an optimized model is created:

- SVM
- Logistic Regression
- Decision Tree
- Random Forest
- XGBoost



03

Results and Recommendation

Best Features

10Days lagged Weather

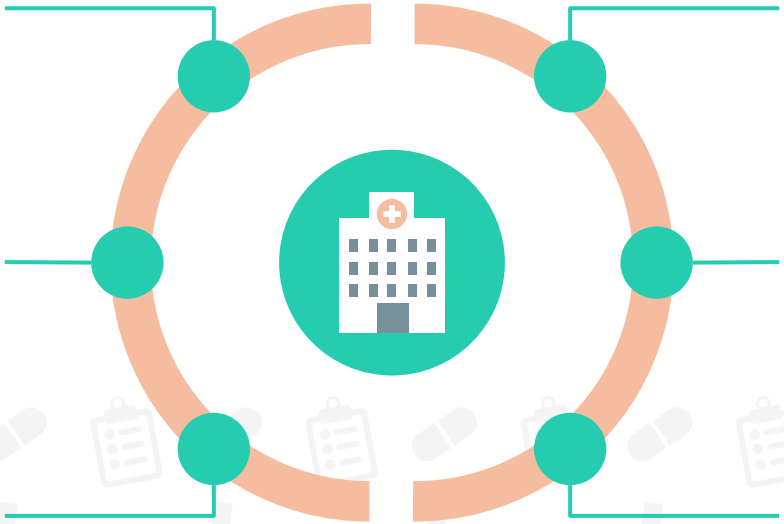
High temperature in previous 10 days leads to WNv spread

Present Weather

High temp/ moderate wind favor the virus spread

Location

Northwestern part of Chicago is the most severe



Temporal features

WNv is widely spread during August

Species

“CULEX PIPIENS”
The trouble maker

Mosquito life cycle

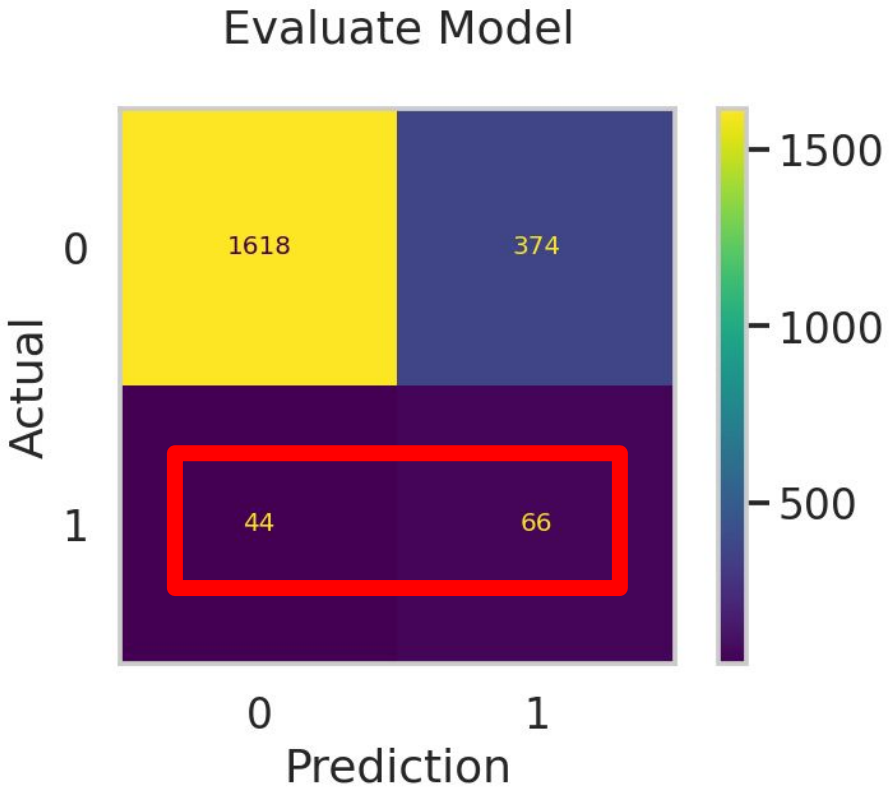
7 days with no rain =
High chance of hatching

Best Model

Random Forest
Kaggle Area under the ROC Curve Score: 0.725

Training Test Section:
(for prediction WNV is present)
AUC Score: 0.718
Precision: 0.15
Recall: 0.63
F1 Score: 0.25

Model's Confusion Matrix



From the best random forest estimator:

- Approximately 60% (66/110) of the WNv positive test result were correctly identified

The Cost

It can approximately cost 701,790 USD per month to spray the whole Chicago. [source]

There are 2.7 million people in Chicago at risk of the infection

- For serve cases it will cost 33,000 USD to treat the Wnv patient
- For non serve cases it can cost up to 7,500 USD to treat the Wnv patient

This mean it can approximately cost up to 89,486,000,000 USD if the whole of Chicago was infected.

The spray cost outweigh the cost of the Wnv treatment, meaning the spray would need to prevent at least 21 cases per month to make it worth the investment. So it is better to implement the spray and prevent a mass outbreak.

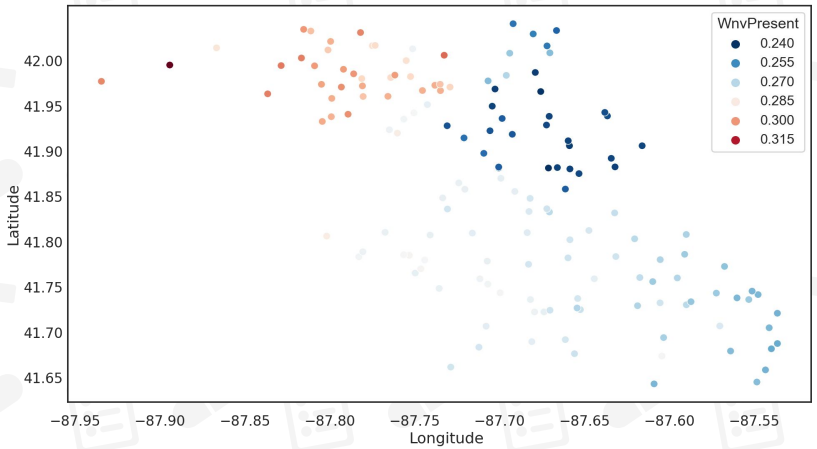
The Probability

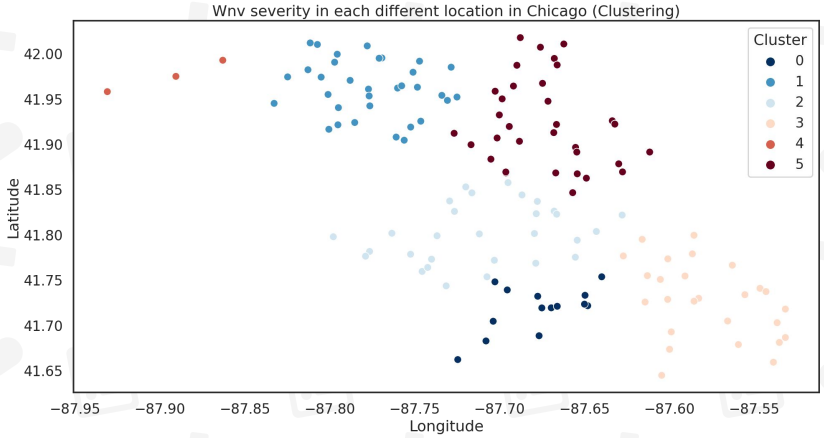
Each cluster represent districts in Chicago:

- 0 - West Englewood
- 1 - East Garfield Park
- 2 - New City
- 3 - South Side
- 4 - Near West Side
- 5 - Central Chicago

The highest probability that Wnv is positive is located in cluster 4

The second highest probability that Wnv is positive is located in cluster 1

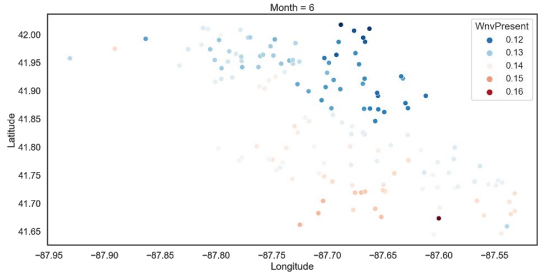


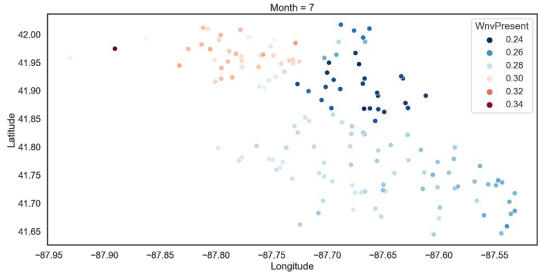


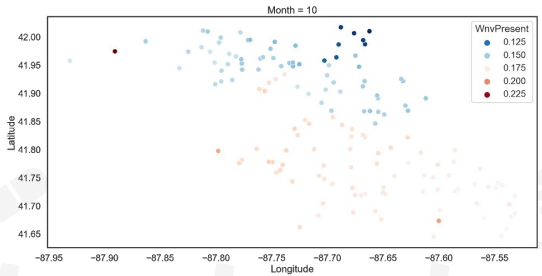
Monthly Data

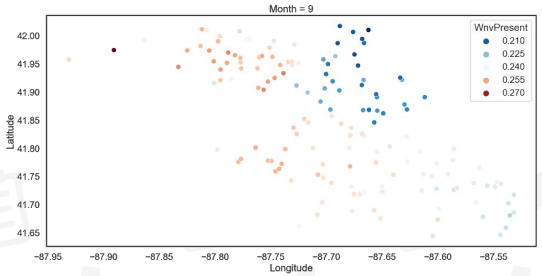
Each month showed different positive Wnv probability in the Chicago District

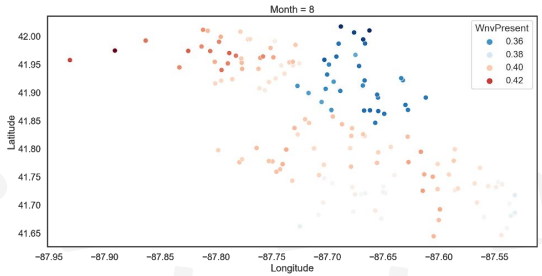
- June - Highest in cluster 0 and 3
- July - Highest in cluster 1 and 5
- August - Highest in cluster 3 and 4
- September - Highest in cluster 2 and 3
- October - Highest in cluster 3 and 4











Cost and Benefit Analysis

In Chicago's spraying program, the city sprays areas when CDPH mosquito traps test positive for WNV two weeks in a row. [source]

However, from our model we can predict with 60% certainty which location and month will be positive without having to wait for two consecutive positive result, this allows the city to:

- Spray with more location and time precision
- Increase the effectiveness of the spray radius
- Reduce number of positive inpatients
- Reduce spraying cost

The ability to predict in advance, and its low implementation cost, will allow our model to be feasible if it could reduce the positive case by only 1 patient

Our Recommendation

The Target

High density area and high Wnv positivity prob cluster*

*varies from month to month

The Opportunity Cost

Might not be able to target other low positive Wnv area

Most Important

Cluster 4 and 1 are most important target*

Least Important

Cluster 2 and 0 are the least affected by Wnv*

04

Conclusion and the Future

Conclusion

The best model used was Random Forest estimators with features like weather patterns, location and mosquitoes life cycle lag.

For the cost and benefit the government and CDC can:

- Target specific location to distribute the Wnv spray
- Consider the cost of the spray and the treatment cost

From the cost and predicted location benefit analysis, the cost of spray are likely to remain the same or decrease since we are able to target more precise location rather than the whole of Chicago which could cost more, while making the spray more effective from the precise location resource allocation.

Additional Data

Mosquitos Numbers

Predicting the mosquitos numbers to use as a feature

Spray Radius

How far does the spray travel and how effective it is

Spray Effectiveness

How effective is the spray in killing mosquitoes

Thank you!